

## THREE DIMENSIONAL STRUCTURE MEMORY

### BACKGROUND OF THE INVENTION

#### 1. Field of the Invention

The present invention relates to stacked integrated circuit memory.

#### 2. State of the Art

Manufacturing methods for increasing the performance and decreasing the cost of electronic circuits, nearly without exception, are methods that increase the integration of the circuit and decrease its physical size per equivalent number of circuit devices such as transistors or capacitors. These methods have produced as of 1996 microprocessors capable of over 100 million operations per second that cost less than \$1,000 and 64Mbit DRAM circuits that access data in less than 50ns and cost less than \$50. The physical size of such circuits is less than 2cm<sup>2</sup>. Such manufacturing methods support to a large degree the economic standard of living in the major industrialized countries and will most certainly continue to have significant consequences in the daily lives of people all over the world.

Circuit manufacturing methods take two primary forms: process integration and assembly integration. Historically the line between these two manufacturing disciplines has been clear, but recently with the rise in the use of MCMs (Multi-Chip Modules) and flip-chip die attach, this clear separation may soon disappear. (The predominate use of the term Integrated Circuit (IC) herein is in reference to an Integrated Circuit in singulated die form as sawed from a circuit substrate such as a semiconductor wafer versus, for example, an Integrated Circuit in packaged form.) The majority of ICs when in initial die form are presently individually packaged, however, there is an increasing use of MCMs. Die in an MCM are normally attached to a circuit substrate in a planar fashion with conventional IC die I/O interconnect bonding methods such as wire bonding, DCA (Direct Chip Attach) or FCA (Flip-Chip Attach).

Integrated circuit memory such as DRAM, SRAM, flash EPROM,

EEPROM, Ferroelectric, GMR (Giant MagnetoResistance), etc. have the common architectural or structural characteristic of being monolithic with the control circuitry integrated on the same die with the memory array circuitry. This established (standard or conventional) architecture or circuit layout structure creates a design trade-off constraint between control circuitry and memory array circuitry for large memory circuits. Reductions in the fabrication geometries of memory cell circuitry has resulted in denser and denser memory ICs, however, these higher memory densities have resulted in more sophisticated control circuitry at the expense of increased area of the IC. Increased IC area means at least higher fabrication costs per IC (fewer ICs per wafer) and lower IC yields (fewer working ICs per wafer), and in the worst case, an IC design that cannot be manufactured due to its non-competitive cost or unreliable operation.

As memory density increases and the individual memory cell size decreases more control circuitry is required. The control circuitry of a memory IC as a percentage of IC area in some cases such as DRAMs approaches or exceeds 40%. One portion of the control circuitry is the sense amp which senses the state, potential or charge of a memory cell in the memory array circuitry during a read operation. The sense amp circuitry is a significant portion of the control circuitry and it is a constant challenge to the IC memory designer to improve sense amp sensitivity in order to sense ever smaller memory cells while preventing the area used by the sense amp from becoming too large.

If this design constraint or trade-off between control and memory circuits did not exist, the control circuitry could be made to perform numerous additional functions, such as sensing multiple storage states per memory cell, faster memory access through larger more sensitive sense amps, caching, refresh, address translation, etc. But this trade-off is the physical and economic reality for memory ICs as they are presently made by all manufacturers.

The capacity of DRAM circuits increase by a factor of four from one gen-

eration to the next; e.g. 1Mbit, 4Mbit, 16Mbit and 64Mbit DRAMs. This four times increase in circuit memory capacity per generation has resulted in larger and larger DRAM circuit areas. Upon introduction of a new DRAM generation the circuit yields are too low and, therefore, not cost effective for high volume manufacture. It is normally several years between the date prototype samples of a new DRAM generation are shown and the date such circuits are in volume production.

Assembling die in a stacked or three dimensional (3D) manner is disclosed in U.S. Patent 5,354,695 of the present inventor, incorporated herein by reference. Furthermore, assembling die in a 3D manner has been attempted with regard to memory. Texas Instruments of Dallas Texas, Irvine Sensors of Costa Mesa California and Cubic Memory Corporation of Scotts Valley California, have all attempted to produce stacked or 3D DRAM products. In all three cases, conventional DRAM circuits in die form were stacked and the interconnect between each DRAM in the stack was formed along the outside surface of the circuit stack. These products have been available for the past several years and have proved to be too expensive for commercial applications, but have found some use in space and military applications due to their small physical size or footprint.

The DRAM circuit type is referred to and often used as an example in this specification, however, this invention is clearly not limited to the DRAM type of circuit. Undoubtedly memory cell types such as EEPROMs (Electrically Erasable Programmable Read Only Memories), flash EPROM, Ferroelectric, or combinations (intra or inter) of such memory cells can also be used with the present Three Dimensional Structure (3DS) methods to form 3DS memory devices.

The present invention furthers, among others, the following objectives:

1. Several-fold lower fabrication cost per megabyte of memory than circuits conventionally made solely with monolithic circuit integration methods.
2. Several-fold higher performance than conventionally made memory circuits.
3. Many-fold higher memory density per IC than conventionally made

memory circuits.

4. Greater designer control of circuit area size, and therefore, cost
5. Circuit dynamic and static self-test of memory cells by an internal controller.
6. Dynamic error recovery and reconfiguration.
7. Multi-level storage per memory cell.
8. Virtual address translation, address windowing, various address functions such as indirect addressing or content addressing, analog circuit functions and various graphics acceleration and microprocessor functions.

#### SUMMARY OF THE INVENTION

The present 3DS memory technology is a stacked or 3D circuit assembly technology. Features include:

1. Physical separation of the memory circuits and the control logic circuit onto different layers;
2. The use of one control logic circuit for several memory circuits;
3. Thinning of the memory circuit to less than about 50 $\mu$ m in thickness forming a substantially flexible substrate with planar processed bond surfaces and bonding the circuit to the circuit stack while still in wafer substrate form; and
4. The use of fine-grain high density inter layer vertical bus connections.

The 3DS memory manufacturing method enables several performance and physical size efficiencies, and is implemented with established semiconductor processing techniques. Using the DRAM circuit as an example, a 64Mbit DRAM made with a 0.25 $\mu$ m process could have a die size of 84mm<sup>2</sup>, a memory area to die size ratio of 40% and a access time of about 50ns for 8Mbytes of storage; a 3DS DRAM IC made with the same 0.25 $\mu$ m process would have a die size of 18.6mm<sup>2</sup>, use 17 DRAM array circuit layers, a memory area to die size ratio of 94.4% and an expected access time of less than 10ns for 64Mbytes of storage. The 3DS DRAM IC manufacturing method represents a scalable, many-fold reduction in the cost per megabyte versus that of conventional DRAM IC manufacturing methods. In other words, the 3DS memory manufacturing method represents, at the infrastruc-

ture level, a fundamental cost savings that is independent of the process fabrication technology used.

#### BRIEF DESCRIPTION OF THE DRAWING

The present invention may be further understood from the following description in conjunction with the appended drawing. In the drawing:

Figure 1a is a pictorial view of a 3DS DRAM IC manufactured with Method A or Method B and demonstrating the same physical appearance of I/O bond pads as a conventional IC die;

Figure 1b is a cross-sectional view of a 3DS memory IC showing the metal bonding interconnect between several thinned circuit layers;

Figure 1c is a pictorial view of a 3DS DRAM IC stack bonded and interconnected face-down onto a larger conventional IC or another 3DS IC;

Figure 2a is a diagram showing the physical layout of a 3DS DRAM array circuit block with one data-line set of bus lines, i.e one port;

Figure 2b is a diagram showing the physical layout of a 3DS DRAM array circuit block with two sets of data-line bus lines, i.e. two ports;

Figure 2c is a diagram showing the physical layout of a portion of an exemplary memory controller circuit;

Figure 3 is a diagram showing the physical layout of a 3DS DRAM array circuit showing partitions for sixty-four (64) 3DS DRAM array blocks;

Figure 4 is a cross-sectional view of a generic 3DS vertical interconnection or feed-through in a thinned substrate;

Figure 5 is a diagram showing the layout of a 3DS memory multiplexer for down-selecting gate-line read or write selection.

#### DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Referring to Figure 1a and Figure 1b, the 3DS (Three Dimensional Structure) memory device 100 is a stack of integrated circuit layers with fine-grain vertical interconnect between all circuit layers. The term fine-grain inter-layer vertical interconnect is used to mean electrical conductors that pass through a circuit layer with or without an intervening device element and have a pitch of nominally less

than 100 $\mu$ m and more typically less than 10 $\mu$ m, but not limited to a pitch of less than 2 $\mu$ m, as best seen in Figure 2a and Figure 2b. The fine-grain inter-layer vertical interconnect also functions to bond together the various circuit layers. As shown in Figure 1b, although the bond and interconnect layers 105a, 105b, etc., are preferably metal, other material may also be used as described more fully hereinafter.

The pattern 107a, 107b, etc. in the bond and interconnect layers 105a, 105b, etc. defines the vertical interconnect contacts between the integrated circuit layers and serves to electrically isolate these contacts from each other and the remaining bond material; this pattern takes the form of either voids or dielectric filled spaces in the bond layers.

The 3DS memory stack is typically organized as a controller circuit 101 and some number of memory array circuit layers 103, typically between nine (9) and thirty-two (32), but there is no particular limit to the number of layers. The controller circuit is of nominal circuit thickness (typically 0.5mm or greater), but each memory array circuit layer is a thinned and substantially flexible circuit with net low stress, less than 50 $\mu$ m and typically less than 10 $\mu$ m in thickness. Conventional I/O bond pads are formed on a final memory array circuit layer for use with conventional packaging methods. Other metal patterns may be used such as insertion interconnection (disclosed in U.S. Patents 5,323,035 and 5,453,404 of the present inventor), DCA (Direct Chip Attach) or FCA (Flip-Chip Attach) methods.

Further, the fine grain inter-layer vertical interconnect can be used for direct singulated die bonding between a 3DS memory die and a conventional die (wherein the conventional die could be the controller circuit as shown in Figure 1c) or a 3DS memory die and another 3DS memory die; it should be assumed that the areas (sizes) of the respective dice to be bonded together can vary and need not be the same. Referring more particularly to Figure 1c, a 3DS DRAM IC stack 100 is bonded and interconnected face-down onto a larger conventional IC or another

3DS IC 107. Optionally the 3DS stack 100 can be composed of only DRAM array circuits with the DRAM controller circuitry as part of the larger die. If the DRAM controller circuitry is part of the larger die, then fine-grain vertical bus interconnect would be required (at the face 109 of the 3DS DRAM IC stack 100) to connect the 3DS DRAM array circuit to the DRAM controller, otherwise larger grain conventional interconnection could be incorporated (patterned) into the planarized bond layer.

As shown in Figure 3, each memory array circuit layer includes a memory array circuit 300 composed of memory array blocks 301 (nominally less than 5mm<sup>2</sup> in area) and each block is composed of memory cells (in much the same manner as the cell array of a DRAM or EEPROM circuit), busing electrodes, and—at the option of the designer—enabling gates for selection of specific rows or columns of the memory array. The controller circuit is composed of sense amps, address, control and drive logic that would normally be found at the periphery of a typical memory circuit of monolithic design such as in a conventional DRAM.

Fine-grain busing vertically connects the controller independently to each memory array layer such that the controller can provide drive (power) or enable signals to any one layer without affecting the state of any of the other layers. This allows the controller to test, read or write independently each of the memory circuit layers.

Figure 2a and Figure 2b show examples of layouts of possible blocks of a memory array such as the block 301 of Figure 3. Although only a portion of the block is shown, in the illustrated embodiment, the blocks exhibit bilateral symmetry such that the layout of the complete block may be ascertained from the illustrated portion. Abbreviations "T", "L", and "TL" are used following various reference numerals to indicate "Top", "Left" and "Top-Left" respectively, implying corresponding elements not shown in the figure.

Referring to Figure 2a, a core portion 200 of the block is composed of a

"sea" of memory cells. Logically, the aggregation of memory cells may be subdivided into "macrocells" 201 each containing some number of memory cells, e.g. an 8 x 8 array of 64 memory cells. At the periphery of the core is formed fine-grain vertical interconnect comprising inter-layer bond and bus contact metallizations 400, described in greater detail hereinafter with reference to Figure 4. The fine-grain vertical interconnect includes I/O power and ground bus lines 203TL, memory circuit layer selects 205T, memory macro cell column selects 207T, data lines 209L, and gate-line multiplexer ("mux") selects 209TL. Gate-line multiplexers 211T are, in the illustrated embodiment, 4:1 multiplexers used to select one of four columns within an eight-wide memory macro cell column. Corresponding bottom-side 4:1 multiplexers combine with the topside multiplexers 211T to form equivalent 8:1 multiplexers for selecting a single gate-line from an eight-gate-line-wide memory macro cell column.

One implementation of a 4:1 gate-line bus multiplexer 500 is shown in Figure 5. Gate-line enables 209TL' (formed in a Metal-1 layer, for example) control transistors 501a through 501d, respectively. Coupled to the transistors are respective gate lines 503a through 503d. Also partly visible are gate-lines 505a through 505d which are coupled to a corresponding 4:1 multiplexer (not shown). When one of the gate-line enables is active, the corresponding gate-line is coupled to an output line 507 of the multiplexer (formed in a Metal-2 layer, for example). The output line is connected to one or more vertical bus connects through a line 509 (formed in a Metal-3 layer and corresponding to metal contact 400 of vertical bus interconnect, for example) and tungsten plugs 511 and 513. The tungsten plug 513 joins the line 509 to a vertical interconnect (not shown).

Referring again to Figure 2a, in the case of a memory circuit layer, the layer may also include output line enables (gates) from controller layer enable signals 205T, for which I/O enables (gates) 213 may be provided.

Note that at the memory layer level, each memory block 301 is electrically

isolated from every other memory block 301. Accordingly, the yield probability for each memory block is independent.

Additional read/write ports can be added as can additional gate-line vertical interconnections; additional vertical interconnection can be used in a redundant manner to improve vertical interconnect yield. The 3DS memory circuit can be designed to have one or more data read and write bus port interconnections. Referring to Figure 2b, a memory block 301' is shown as having a port  $P_0$ , (209L) and a further port  $P_1$  (209L'). The only limitation on the number of vertical interconnections is the overhead such vertical interconnections impose on the cost of the circuit. The fine-grain vertical interconnect method allows thousands of interconnects per block at an increase in die area of only a few percent.

As an example, the overhead of the vertical interconnect shown in Figure 2b for a DRAM memory block of 4Mbits with two read/write ports and implemented in  $0.35\mu m$  or  $0.15\mu m$  design rules consists of approximately 5,000 connections and is less than 6% of the total area of the memory array block. Therefore, the vertical interconnect overhead for each memory array circuit layer in the 3DS DRAM circuit is less than 6%. This is significantly less than that presently experienced in monolithic DRAM circuit designs where the percentage of non-memory cell area can exceed 40%. In a completed 3DS DRAM circuit the percentage of non-memory cell area is typically less than 10% of the total area of all circuits in the stacked structure.

The 3DS memory device decouples control functions that normally would be found adjacent the memory cells of monolithic memory circuits and segregates them to the controller circuit. The control functions, rather than occurring on each memory array layer as in conventional memory ICs, occur only once in the controller circuit. This creates an economy by which several memory array layers share the same controller logic, and therefore, lowers the net cost per memory cell by as much as a factor of two versus conventional memory design.

The segregation of the control functions to a separate controller circuit allows more area for such functions (i.e., an area equal to the area one or several of the memory array blocks). This physical segregation by function also allows fabrication process segregation of the two very different fabrication technologies used for the control logic and the memory array, again realizing additional fabrication cost savings versus the more complicated combined logic/memory fabrication process used for conventional memory. The memory array can also be fabricated in a process technology without consideration of the process requirements of control logic functions. This results in the ability to design higher performance controller functions at lower cost than is the case with present memory circuits. Furthermore, the memory array circuit can also be fabricated with fewer process steps and nominally reduce memory circuit fabrication costs by 30% to 40% (e.g., in the case of a DRAM array, the process technology can be limited to NMOS or PMOS transistors versus CMOS).

Hence, although bonding of sufficiently planar surfaces of a memory controller substrate and a memory array substrate using thermal diffusion metal bonding is preferred, in the broader aspects of the present invention, the invention contemplates bonding of separate memory controller and memory array substrates by any of various conventional surface bonding methods, such as anisotropically conductive epoxy adhesive, to form interconnects between the two to provide random access data storage.

Referring to Figure 2c, the layout of a portion of an exemplary memory controller circuit is shown. The inter-layer bond and bus contact metallization has the same pattern as previously described in relation to Figure 2a. Instead of a sea of memory cells, however, there is provided memory controller circuitry including, for example, sense amps and data line buffers 215. Because of the increased availability of die area, multi-level logic may be provided in conjunction with the sense amps and data line buffers 215. Also shown are address decode, gate-line and

DRAM layer select logic 217, refresh and self-test logic 219, ECC logic 221, windowing logic 223, etc. Note that self-test logic, ECC logic, and windowing logic are provided in addition to functions normally found within a DRAM memory controller circuit. Depending on die size or the number of controller circuit layers used, any of numerous other functions may also be provided including, for example, virtual memory management, address functions such as indirect addressing or content addressing, data compression, data decompression, audio encoding, audio decoding, video encoding, video decoding, voice recognition, handwriting recognition, power management, database processing, graphics acceleration functions, microprocessor functions (including adding a microprocessor substrate), etc.

The size of the 3DS memory circuit die is not dependent on the present constraint of containing the necessary number of memory cells and control function logic on one monolithic layer. This allows the circuit designer to reduce the 3DS circuit die size or choose a die size that is more optimal for the yield of the circuit. 3DS memory circuit die size is primarily a function of the size and number of memory array blocks and the number of memory array layers used to fabricate the final 3DS memory circuit. (The yield of a nineteen (19) layer,  $0.25\mu\text{m}$  process 3DS DRAM memory circuit may be shown to be greater than 90% as described below.) This advantage of selecting the 3DS circuit die size enables an earlier first production use of a more advanced process technology than would normally be possible for conventional monolithic circuit designs. This, of course, implies additional cost reductions and greater performance over the conventional memory circuits.

### **3DS Memory Device Fabrication Methods**

There are two principal fabrication methods for 3DS memory circuits. The two 3DS memory fabrication methods, however, have a common objective which is the thermal diffusion metal bonding (also referred to as thermal compression

bonding) of a number of circuit substrates onto a rigid supporting or common substrate which itself may optionally also be a circuit component layer.

The supporting or common substrate can be a standard semiconductor wafer, a quartz wafer or a substrate of any material composition that is compatible with the processing steps of the 3DS circuit, the operation of the circuit and the processing equipment used. The size and shape of the supporting substrate is a choice that best optimizes available manufacturing equipment and methods. Circuit substrates are bonded to the supporting substrate and then thinned through various methods. Circuit substrates may be formed on standard single crystal semiconductor substrates or as polysilicon circuits formed on an appropriate substrate such as silicon or quartz.

Polysilicon transistor circuits have the important cost saving option of incorporating a parting layer (film) that allows the substrate upon which the polysilicon circuits are formed to be released and reused. Polysilicon transistor or TFTs (Thin Film Transistor) devices are widely used, and need not be made solely from silicon.

The various circuit layers of the 3DS memory circuit are bonded together by use of thermal diffusion of two metal surfaces, typically aluminum. The surfaces of the circuits to be bonded are smooth and sufficiently planar as is the case with the surface of an unprocessed semiconductor wafer or a processed semiconductor wafer that has been planarized with the CMP (Chemical Mechanical Processing) method with a surface planarity of less than  $1\mu\text{m}$  and preferably less than  $1,000\text{\AA}$  over at least the area of the surface of the circuit (formed on the substrate) to be bonded. The metal bonding material on the surfaces of the circuits to be bonded are patterned to be mirror images of each other and to define the various vertical interconnect contacts as indicated in Figure 2a, Figure 2b, Figure 2c and Figure 5. The step of bonding two circuit substrates results in simultaneously forming the vertical interconnection between the two respective circuit layers or substrates.

The thermal diffusion bonding of the circuit layers takes place preferably in an equipment chamber with controlled pressure and atmospheric components such as N<sub>2</sub> with little H<sub>2</sub>O and O<sub>2</sub> content. The bonding equipment aligns the patterns of the substrates to be bonded, presses them together with a set of programmed pressures and at one or more temperatures for a period of time as required by the type of metal used as the bonding material. The thickness of the bonding material is nominally in a range of 500Å to 15,000Å or greater with a preferred thickness of 1,500Å. The initial bonding of the substrates is preferably done at lower than standard pressure such as a negative pressure between 1 torr and 740 torr depending on the design of the bond pattern. This can leave an interior negative pressure between the bonding surfaces once external atmospheric pressure is returned which further assists in the formation of the bond and enhances the reliability of the bond.

The preferred bonding material is pure aluminum or an alloy of aluminum, but it is not limited to aluminum and may include, for example, such metals as Sn, Ti, In, Pb, Zn, Ni, Cu, Pt, Au or alloys of such metals that provide acceptable surface bond diffusion capabilities at acceptable temperatures and forming periods. The bonding material is not limited to metal, and could be a combination of bonding materials, such as highly conductive polysilicon, some of which are non-conducting such as silicon dioxide, and the foregoing exemplary types of bond material choices should not be considered to be limitations on how the circuit layers can be bonded.

In the case where metal bond materials form a native surface oxide that either inhibits the forming of a satisfactory bond or may increase the resistance in the vertical interconnections formed by the bond, the oxide should be removed. The bonding equipment provides an oxide reduction capability such that bonding surfaces of the bond material are rendered without native surface oxide. The methods of forming gas atmospheres for the reduction of surface oxides are well

known, and there are other methods for removing the native oxide such as sputter etching, plasma etching or ion mill etching. In the case where aluminum is used as the bonding material, it is preferred that the thin native aluminum oxide film of approximately 40Å on the bonding surfaces be removed prior to bonding.

The thinned (substantially flexible) substrate circuit layers of the 3DS memory circuit are typically memory array circuits, however, the thinned substrate circuit layers are not limited to memory circuits. Other circuit layer types can be controller circuits, non-volatile memory such as EEPROM, additional logic circuitry including microprocessor logic and application specific logic functions such as those that support graphic or database processing, etc. The selection of such circuit layer types follows from the functional requirements of the design of the circuit and is not limited by the 3DS memory fabrication process.

The thinned (substantially flexible) substrate circuit layers are preferably made with dielectrics in low stress (less than  $5 \times 10^8$  dynes/cm<sup>2</sup>) such as low stress silicon dioxide and silicon nitride dielectrics as opposed to the more commonly used higher stress dielectrics of silicon oxide and silicon nitride used in conventional memory circuit fabrication. Such low stress dielectrics are discussed at length in U.S. Patent 5,354,695 of the present inventor, incorporated herein by reference. The use of dielectrics with conventional stress levels could be used in the assembly of a 3DS DRAM circuit, however, if more than a few layers comprise the stacked assembly, each layer in the assembly will have to be stress balanced so that the net stress of the deposited films of a layer is less than  $5 \times 10^8$  dynes/cm<sup>2</sup>. The use of intrinsically low stress deposited films is the preferred method of fabrication versus the use of the method where the stress of individually deposited films are not equal but are deposited to create a net balanced lower stress.

#### **Method A, 3DS memory device fabrication sequence**

This fabrication sequence assumes that several circuit layers will be bonded to a common or support substrate and subsequently thinned in place. An

example of a resulting 3DS memory circuit is shown in Figure 1a.

1. Align and bond to the common substrate the topside of a second circuit substrate.

2A. Grind the backside or exposed surface of the second circuit substrate to a thickness of less than 50 $\mu$ m and then polish or smooth the surface. The thinned substrate is now a substantially flexible substrate.

Optionally an etch stop may be incorporated in the second substrate from less than a micron to several microns below the semiconductor surface prior to device fabrication. This etch stop can be an epitaxially formed film such as GeB (described in U.S. Patents 5,354,695 and 5,323,035 of the present inventor, incorporated herein by reference) or a low density implanted layer of O<sub>2</sub> or N<sub>2</sub> to form a buried oxide or nitride barrier etch stop layer just below the device layer on the topside of the second substrate. After a preliminary grinding of a significant portion of the backside of the substrate, the remaining portion of the backside of the second substrate is then selectively etched in a chemical bath which stops on the surface of the epitaxial or implanted layer. Subsequent polishing and RIE steps as necessary can then be used to complete the thinning of the second substrate.

Alternately, a parting layer such as H<sub>2</sub> implanted into the topside surface of the second substrate prior to device fabrication can be used with a thermal step to crack off the majority of the backside of the second substrate, allowing its reuse.

2B. The second substrate may alternatively be a circuit formed of polysilicon transistors or TFTs over a parting layer such as aluminum, titanium, AlAs, KBr, etc. which can be activated by a specific chemical release agent. The backside of the second substrate is then removed upon activating (dissolving) the release layer and followed as needed by interconnect semiconductor processing steps.

3. Process the thinned backside of the second substrate to form vertical interconnections such as that shown in Figure 4 with the bonded surface side of the second substrate. The backside processing typically comprises conventional semiconductor processing steps of dielectric and metal deposition, lithography

and RIE, the order of which can vary to a great degree. The completion of the backside processing will also result in a patterned metal layer that is similar to the topside bond material pattern to facilitate the subsequent bonding of an additional circuit substrate, a terminal pattern such as a conventional I/O IC bond pad (wire bonding) pattern, a pattern for thermal diffusion bonding of the 3DS memory circuit to another die (either another 3DS circuit or a conventional die), or a pattern for insertion interconnection, conventional DCA (Direct Chip Attach) or FCA (Flip-Chip Attach).

Referring more particularly to Figure 4, during the fabrication of active circuit devices, an oxide mask 401 is thermally grown or deposited. Vertical bus contacts 403 are then formed, for example from highly-doped polysilicon coincident with a polysilicon gate forming step. Alternatively, contact 403 may be formed of metal. Conventional DRAM interconnect structures 410 are then formed using conventional processing. The DRAM interconnect may include an internal pad 405. The "DRAM processed" portion 420 of the wafer includes various dielectric and metal layers. A final passivation layer 407 is deposited, after which vias 409 are formed. Conventional CMP processing is then used to obtain a planar surface 411. Contacts 413 and bond surfaces not shown are then patterned in a top-most metal layer (e.g. Metal-3).

After bonding and thinning of the backside of the second substrate to about 1-8 $\mu$ m of silicon (or other semiconductor) substrate 415, feed-throughs 417 are then formed in registration with the contacts 403. A passivation layer 419 and contacts 421 are then formed. The contacts 421 may be formed so as to form a mirror image of the contacts 413, allowing for the bonding of further wafers.

4. If another circuit layer is to be bonded to the 3DS circuit stack, steps 1-3 are repeated.

5A. The circuits of the finished 3DS memory substrate are then conventionally sawed into die (singulated), resulting in a circuit of the type shown in Figure 1a, and packaged as would be the case with conventional integrated circuits.

5B. The circuits of the finished 3DS memory substrate are then conventionally sawed and then individually aligned and thermal diffusion bonded (metal pattern down) to the surface of a second (conventional IC) die or MCM substrate in a manner similar to that used in the bonding of the circuit substrates of step 1 above. (The conventional die or MCM substrate may have a larger area than the 3DS memory substrate and may include a graphics controller, video controller or microprocessor, such that the 3DS becomes embedded as part of another circuit.) This final bonding step typically incorporates a fine-grain interconnect between the 3DS memory circuit and the die or MCM substrate, but could also use a conventional interconnect pattern. Further, a 3DS memory circuit can be bonded face up to a conventional IC in die form or MCM substrate and wire bonding used to form conventional I/O interconnections.

#### Method B, 3DS memory device fabrication sequence

This fabrication sequence assumes that a circuit substrate will first be bonded to a transfer substrate, thinned and then bonded to a common substrate as a layer of the circuit stack. The transfer substrate is then released. This method has the advantage over Method A of allowing substrates to be thinned prior to being bonded to the final circuit stack and allows for simultaneous thinning and vertical interconnect processing of substrate circuit layers.

1. Bond to a transfer substrate a second circuit substrate using a release or parting layer. A transfer substrate may have high tolerance parallel surfaces (TTV or Total Thickness Variance of less than  $1\mu\text{m}$ ) and may be perforated with an array of small holes to assist the parting process. The parting layer can be a blanket deposition of a bonding metal. Precise alignment of the surfaces is not required.

2. Perform step 2A or 2B of Method A.

3. Process the backside of the second substrate to form interconnections with the bonded topside surface of the second substrate as shown in Figure 4. The backside processing typically comprises conventional semiconductor processing steps of dielectric and metal deposition, lithography and RIE, the order of which can vary to great degree. The completion of the backside processing will also

result in a patterned metal layer that is similar to the bond material pattern of the common substrate to facilitate the subsequent bonding of an additional circuit layer.

4. Bond the second circuit to a common or support substrate (3DS stack) and release the transfer substrate by activating the parting layer between it and the second circuit.

5. Process the now exposed topside of the second substrate to form interconnections for subsequent substrate bonding or a terminal pattern for conventional I/O bonding (wire bonding) pad pattern, a pattern for thermal diffusion bonding of the 3DS memory circuit to another die (either another 3DS circuit or a conventional die), or a pattern for conventional insertion interconnect, DCA (Direct Chip Attach) or FCA (Flip-Chip Attach). If another circuit layer is to be bonded to the 3DS circuit stack, steps 1 through 4 are repeated.

6. Perform step 5A or 5B of Method A.

### **3DS Memory Device Yield Enhancement Methods**

The 3DS circuit may be considered a vertically assembled MCM (Multi-Chip Module) and as with an MCM the final yield is the product of the yield probabilities of each component circuit (layer) in the completed 3DS circuit. The 3DS circuit uses several yield enhancement methods that are synergistic in their combined usage within a single memory IC. The yield enhancement methods used in the 3DS memory circuit include small memory array block size, memory array block electrical isolation through physically unique or separate vertical bus interconnections, intra memory array block gate-line sparing, memory array layer sparing (inter-block gate-line sparing), controller sparing and ECC (Error Correcting Codes). The term sparing is used to mean substitution by a redundant element.

The selected size of the memory array block is the first component in the yield equation for the 3DS memory circuit. Each memory array block is individually (uniquely) accessed and powered by the controller circuit and is physically independent of each and every other memory array block including those on the

same memory array layer in addition to those on a different memory array layer. The size of the memory array block is typically less than 5mm<sup>2</sup> and preferably less than 3mm<sup>2</sup>, but is not limited to a specific size. The size of memory array block, the simplicity of its NMOS or PMOS fabrication process and its physical independence from each of the other memory array blocks, for nearly all production IC processes, provides a conservatively stated nominal yield of greater than 99.5%. This yield assumes that most point defects in the memory array block such as open or shorted interconnect lines or failed memory cells can be spared (replaced) from the intra-block or inter-block set of redundant gate-lines. Major defects in a memory array block which render the complete memory array block unusable result in the complete sparing of the block from a redundant memory array layer or the rejection of the 3DS circuit.

In the example of a 3DS DRAM circuit the yield of a stack of memory array blocks is calculated from the yield equation  $Y_s = ((1-(1-P_y)^2)^n)^b$ , where  $n$  is the number DRAM array layers,  $b$  is the number of blocks per DRAM array and  $P_y$  is the effective yield (probability) of a DRAM array block less than 3mm<sup>2</sup> in area. Assuming a DRAM array block redundancy of 4% for gate-lines in the DRAM array block lines and one redundant DRAM array layer, and assuming further that the number of blocks per layer is 64, the number of memory array layers in the stack is 17 and the effective value for  $P_y$  is 0.995, then the stack yield  $Y_s$  for the complete memory array (including all memory array block stacks) is 97.47%.

The  $Y_s$  memory array stack yield is then multiplied by the yield of the controller  $Y_c$ . Assuming a die size of less than 50mm<sup>2</sup>, a reasonable  $Y_c$  for a controller fabricated from a 0.5μm BiCMOS or mixed signal process would be between 65% and 85%, giving a net 3DS memory circuit yield of between 63.4% and 82.8%. If a redundant controller circuit layer is added to the 3DS memory stack, the yield probabilities would be between 85.7% and 95.2%.

The effective yield of a memory array block can be further increased by the

optional use of ECC logic. ECC logic corrects data bit errors for some group size of data bits. The syndrome bits necessary for the operation of ECC logic would be stored on redundant gate-lines of any of the memory array layers in a vertically associated block stack. Further, if necessary, in order to accommodate the storage of ECC syndrome bits, additional memory array layers could be added to the circuit.

#### **Advantageous 3DS Memory Device Controller Capabilities**

As compared to a conventional memory circuit, the 3DS memory controller circuit can have various advantageous capabilities due the additional area available for controller circuitry and the availability of various mixed signal process fabrication technologies. Some of these capabilities are self-test of memory cells with dynamic gate-line address assignment, virtual address translation, programmable address windowing or mapping, ECC, data compression and multi-level storage.

Dynamic gate-line address assignment is the use of programmable gates to enable the layer and gate-line for a read/write operation. This allows the physical order of memory storage to be separate or different from the logical order of stored memory.

The testing of each generation of memory devices has resulted in significantly increased test costs. The 3DS memory controller reduces the cost of testing by incorporating sufficient control logic to perform an internal test (self-test) of the various memory array blocks. Circuit testing in the conventional ATE manner is required only for verification of controller circuit functions. The scope of the internal test is further extended to the programmable (dynamic) assignment of unique addresses corresponding to the various gate-lines of each memory array block on each layer. Self-test capability of the 3DS controller circuit can be used anytime during the life of the 3DS memory circuit as a diagnostic tool and as a means to

increase circuit reliability by reconfiguring (sparing) the addresses of gate-lines that fail after the 3DS memory circuit is in use in a product.

ECC is a circuit capability that, if included in the controller circuit, can be enabled or disabled by a programming signal or made a dedicated function.

Data compression logic will allow the total amount of data that can be stored in the 3DS memory array to be increased. There are various generally known data compression methods available for this purpose.

Larger sense amps allow greater dynamic performance and enable higher speed read operations from the memory cells. Larger sense amps are expected to provide the capability to store more than one bit (multi-level storage) of information in each memory cell; this capability has already been demonstrated in non-volatile memory circuits such as flash EPROM. Multi-level storage has also been proposed for use in the 4Gbit DRAM generation circuits.

It will be appreciated by those of ordinary skill in the art that the invention can be embodied in other specific forms without departing from the spirit or essential character thereof. The presently disclosed embodiments are therefore considered in all respects to be illustrative and not restrictive. The scope of the invention is indicated by the appended claims rather than the foregoing description, and all changes which come within the meaning and range of equivalents thereof are intended to be embraced therein.